2

The Space of Cognitive Vision

David Vernon

DIST, University of Genova, Italy
vernon@ieee.org

Abstract. Cognitive vision is an area that is not yet well-defined, in the sense that one can unambiguously state what issues fall under its purview and what considerations do not. Neither is there unequivocal consensus on the right approach to take in addressing these issues — there isn't a definitive universally-accepted scientific theory with 'gaps in understanding' that merely need to be plugged. On the contrary, there are clearly competing viewpoints and many poorly-understood issues (such as the point where vision stops and cognition starts). Depending on how you choose to view or define cognitive vision, there are many points of departure, some based squarely in artificial intelligence and image processing, others in developmental psychology and cognitive neuroscience, and others yet in cognitive robotics and autonomous systems theory. This paper is an attempt to sketch a framework within which the complete domain of cognitive vision can be set, a framework that embraces all of the possible approaches that can be taken and that highlights common concerns as well as fundamental differences between the approaches. Our goal here is to define cognitive vision in a way that avoids alienating any particular community and to state what the options are. While we will note in passing possible strengths and weaknesses of the various approaches, this paper will not attempt to argue in favour of one approach over another.

2.1 The Background to Cognitive Vision

It is nearly forty years since Roberts first published the results of his seminal attempts to construct a computer vision system [374]. Since then, computer vision has matured and undergone many stages in its evolution. From the blocks-world approaches of the sixties and early seventies [164, 201, 483, 419], to the knowledge-based and model-based approaches of the mid to late seventies [23, 171, 446, 54], the modular information processing approaches of the late seventies and early eighties with their strong emphasis on early vision [278, 30, 280, 283, 284, 282, 193, 281], the development of appearance-based vision in the nineties [81] – a decade that was perhaps distinguished more than anything by the creation of mathematically-sound robust early vision and the associated expansion of vision based on computational geometry [117, 175] – to the more recent probabilistic techniques and the increasingly-widespread use of machine learning [355]. On the way, computer vision has spawned a number of successful offshoots, such as machine vision for industrial inspection, the analysis of video data for remote monitoring of events, and the use of image analysis in the creation of special effects in the film

H.I. Christensen and H.-H. Nagel (Eds.): Cognitive Vision Systems, LNCS 3948, pp. 7–24, 2006.

industry. However, to date, the ultimate goal of creating a general-purpose vision system with anything close to the robustness and resilience of the human visual system remains as elusive as ever.

One of the more recent trends in computer vision research in the pursuit of human-like capability is the coupling of cognition and vision into cognitive computer vision. Unfortunately, it is apparent that the term cognitive computer vision means very different things to different people. For some, it means the explicit use of knowledge and reasoning together with sensory abstraction of data from a perceived environment; for others it implies the emergent behaviour of a physically-active system that learns to make perceptual sense of its environment as it interacts within that environment and as a consequence of that interaction. For others yet, it is a meaningless term in its own right and cannot be treated except as an intrinsic component of the process of cognition that, in turn, is an inherent feature of autonomous systems. Our goal here is to present all of these viewpoints in a single consistent framework:

1. To provide a definition of cognitive vision that is neutral with respect to possible approaches and to explain what capabilities might be provided by such a system;
2. To delineate the space of cognitive vision and characterize it in terms or dimensions that allow it to be mapped on to different approaches;
3. To highlight contentious and significant issues (*e.g.* the necessity for embodiment, the nature and need for representations, the nature and role of knowledge, the role of language, the inter-dependence of perception and action).

These are the issues to which we now turn.

2.2 Towards a Universal Definition of Cognitive Vision

There are several ways one can approach the definition of a discipline. One can take a functional approach, setting out the minimal tasks that a system should be able to carry out, or one can take an architectural approach, identifying the manner in which a system should be constructed and the functional modules that would be used in a typical system. Alternatively, one can adopt a behavioural but non-functional approach that identifies generic attributes, capabilities, and characteristics. A good definition should be neutral to any underlying model, otherwise it begs the research question and preempts the research agenda. Consequently, this rules out an architectural definition. A good definition should also be application-independent. This rules out a strictly functional definition, or at the very least necessitates that any functions be generic and typically common to all possible systems. Consequently, we will attempt to define cognitive vision using generic functionality (*i.e. capability*) and non-functional attributes.

We'll begin with the definition adopted by *ECVision* to date [12]:

> "Cognitive computer vision is concerned with integration and control of vision systems using explicit but not necessarily symbolic models of context, situation and goal-directed behaviour. Cognitive vision implies functionalities for knowledge representation, learning, reasoning about events & structures, recognition and categorization, and goal specification, all of which are concerned with the semantics of the relationship between the visual agent and its environment.'

Although this definition is useful, in that it focusses on many of the key issues, it depends a little too much on architectural issues (*e.g.* integration, control, functional modules) and it is not as neutral to underlying model(s) as perhaps it should be. That is, it strays from a definition of *what* cognitive vision is to a definition of *how* it is to be achieved. As we will see in Section 2.3, there are several competing approaches, not all of which are compatible with the one that is implicitly favoured in this definition. That said, however, it does provide us with a good starting point and the following is an attempt both to expand on it, drawing out the key issues even more, eliminating the model-dependent and architecture-specific components, and highlighting the generic functionalities and non-functional attributes.

> A cognitive vision system can achieve the four levels of generic functionality of a computer vision system:[1]
>
> 1. *Detection* of an object or event in the visual field;
> 2. *Localization* of the position and extent of a detected entity;
> 3. *Recognition* of a localized entity by a labelling process;
> 4. *Understanding* or comprehending the role, context, and purpose of a recognized entity.[2]
>
> It can engage in purposive goal-directed behaviour, adapting to unforeseen changes of the visual environment, and it can anticipate the occurrence of objects or events. It achieves these capabilities through:
>
> 1. a faculty for learning semantic knowledge (*i.e.* contextualized understanding of form and function), and for the development of perceptual strategies and behaviours;
> 2. the retention of knowledge about the environment, the cognitive system itself, and the relationship between the system and its environment;[3]
> 3. deliberation about objects and events in the environment, including the cognitive system itself.

This definition focusses on what constitutes a cognitive vision system, how it should behave, what it should be capable of achieving, and what are its primary characteristics. The first four points encapsulate generic functionality. The next set of issues deal with non-functional attributes, and the final three points suggest a way of spanning the space of cognitive vision.

The three non-functional characteristics of purposive behaviour, adaptability, and anticipation, taken together, allow a cognitive vision system to achieve certain goals, even in circumstances that were not expected when the system was being designed. This capacity for plastic resilient behaviour is one of the hallmarks of a cognitive vision system. The characteristic of anticipation is important as it requires the system to operate

[1] These four levels were suggested by John Tsotsos, York University, during the course of Dagstuhl Seminar 03441[73].

[2] Implicit in the fourth level is the concept of categorization: the assignment of an object or event to a meta-level class on some basis other than visual appearance alone.

[3] The distinction between environmental states, system states, and the environment-system relationship was introduced by Hans-Hellmut Nagel, Universität Karlsruhe, during the course of Dagstuhl Seminar 03441[73].

across a variety of time-scales, extending into the future, so that it is capable of more than reflexive stimulus-response behaviour.

The final three characteristics of cognitive vision — learning, memory, and deliberation — are all concerned with knowledge: its acquisition, storage, and usage. Knowledge is the key to cognitive vision. These three issues highlight the chief differentiating characteristics of cognitive vision *vis-à-vis* computer vision and, as we will see, allow us to define the space of cognitive vision in a way that is relevant to all the various approaches.

First, however, we must survey the different paradigms or approaches that attempt to model and effect these characteristics of cognitive vision.

2.3 A Review of Approaches to Cognition

If we are to understand in a comprehensive way what is meant by cognitive vision, we must address the issue of cognition. Unfortunately, there is no universally-accepted agreement on what cognition is and different research communities have fundamentally different perspectives on the matter.

Broadly speaking, we can identify two distinct approaches to cognition, each of which makes significantly different assumptions about the nature of cognition, the purpose or function of cognition, and the manner in which cognition is achieved. These are:

1. the *cognitivist* approach based on information processing symbolic representational systems;
2. the *emergent systems* approach, embracing connectionist systems, dynamical systems, and enactive systems.

Cognitivist approaches correspond to the classical and still prevalent view that 'cognition is a type of computation' which operates on symbolic representations, and that cognitive systems 'instantiate such representations physically as cognitive codes and ... their behaviour is a causal consequence of operations carried out on these codes' [360]. Connectionist, dynamical, and enactive systems can be grouped together under the general heading of emergent systems that, in contradistinction to the cognitivist view, argues against the information processing view of cognition as 'symbolic, rational, encapsulated, structured, and algorithmic', and argues in favour of one that treats cognition as emergent, self-organizing, and dynamical [447, 219].

2.3.1 Symbolic Information Processing Representational Cognitivist Models

Cognitive science has its origins in cybernetics (1943-53), following the first attempts to formalize what had to that point been metaphysical treatments of cognition. The intention of the early cyberneticians was to create a science of mind, based on logic. Examples of progenitors include McCulloch and Pitts and their seminal paper 'A logical calculus immanent in nervous activity' [294]. This initial wave in the development of a science of cognition was followed in 1956 by the development of an approach

referred to as *cognitivism*. Cognitivism asserts that cognition can be defined as computations on symbolic representations, *i.e.* cognition is information processing [471]: rule-based manipulation of symbols. Much of artificial intelligence research is carried out on the assumption of the correctness of the cognitivist hypothesis. Its counterpart in the study of natural biologically-implemented (*e.g.* human) cognitive systems is cognitive psychology which uses 'computationally characterizable representations' as the main explanatory tool [471]. The entire discipline of cognitive science is often identified with this particular approach [219]: that cognition involves computations defined over internal representations *qua* knowledge, in a process whereby information about the world is abstracted by perception, and represented using some appropriate symbolic data-structure, reasoned about, and then used to plan and act in the world. This approach has also been labelled by many as the *information processing* (or symbol manipulation) approach to cognition [279, 176, 352, 222, 447, 219] whereby perception and cognition entail computations defined over explicit internal representations. It is undoubtedly the prevalent approach to cognition today but, as we will see, it is by no means the only paradigm in cognitive science.

For cognitivist systems, cognition is representational in a strong and particular sense: it entails the manipulation of explicit symbolic representations of the state and behaviour of the external world to facilitate appropriate, adaptive, anticipatory, and effective interaction, and the storage of the knowledge gained from this experience to reason even more effectively in the future. Vision and perception are concerned with the abstraction of isomorphic spatio-temporal representations of the external world from sensory data. Reasoning itself is symbolic: a procedural process whereby explicit representations of an external world are manipulated to infer likely changes in the configuration of the world (and attendant perception of that altered configuration) arising from causal actions.

The term *representation* is heavily laden with meaning. As Davis *et al.* noted [93], knowledge representations entail several issues. First, that the representation is a 'surrogate' or substitute for a thing itself; second, it involves a set of ontological commitments and therefore defines the way we think of or conceive of the world. Third, it is a 'fragmentary theory of intelligent reasoning' and a medium for efficient computation. Finally, representations provide a medium for human expression.

Since the representations deployed in a cognitivist system are the idealized (descriptive) product of a human designer, they can be directly accessed and understood or interpreted by a human and there is no requirement for embodiment to facilitate the development of semantic knowledge. However, one can argue that this is also the key limiting factor of cognitivist vision systems: these programmer-dependent representations effectively bias the system (or 'blind' the system [494]) and constrain it to an idealized description that is dependent on and a consequence of the cognitive requirements of human activity. This works as long as the system doesn't have to stray too far from the conditions under which these observations were made. The further one does stray, the larger the 'semantic gap' [429] between perception and possible interpretation, a gap that is normally plugged by the embedding of (even more) programmer knowledge or the enforcement of expectation-driven constraints [340] to render a system practicable in a given space of problems.

It is easy to see how this approach usually then goes hand-in-hand with the fundamental assumption that 'the world we perceive is isomorphic with our perceptions of it as a geometric environment' [416]. The goal of cognition, for a cognitivist, is to reason symbolically about these representations in order to effect intelligent, adaptive, anticipatory, goal-directed, behaviour. Typically, this approach to cognition will deploy an arsenal of techniques including machine learning, probabilistic modelling, and other techniques intended to deal with the inherent uncertain, time-varying, and incomplete nature of the sensory data that is being used to drive this representational framework.

2.3.2 Emergent Approaches

The next phase in the development of cognitive science was based on emergent systems: systems epitomized by connectionist and dynamical approaches in which cognition is viewed as an emergent property of a network of component elements. Emergence is presented as an alternative to symbol manipulation.

From the perspective of emergent systems, cognition is a process of self-organization whereby the system is continually re-constituting itself in real-time to maintain its operational identity through moderation of mutual system-environment interaction and co-determination. Vision and perception are concerned with the acquisition of visual sensory data to enable effective action and, significantly, are functionally dependent on the richness of the action interface.

Connectionist Models

Connectionist models had their genesis in the study of artificial neural networks and self-organizing systems, *e.g.* the perceptron built by F. Rosenblatt in 1958. However, it lay dormant for some 30 years until a rekindling of interest in artificial neural networks in the late '70s and '80s in the work of Grossberg, Linsker, and many others. One of the motivations for work on emergent systems was disaffection with the sequential and localized character of symbol-manipulation based cognitivism. Emergent systems, on the other hand, depend on parallel and distributed architectures which are biologically more plausible. One of the key features of emergent systems, in general, and connectionism, in particular, is that 'the system's connectivity becomes inseparable *from its history of transformations*, and related to the kind of task defined for the system' [471]. Furthermore, symbols play no role.[4] Whereas in the cognitivist approach the symbols are distinct from what they stand for, in the connectionist approach, meaning relates to the global state of the system and actually the meaning is something attributed by an external third-party observer to the correspondence of a system state with that of the world in which the emergent system is embedded.

Connectionist approaches are for the most part associative learning systems in which the learning phase is either unsupervised (self-organizing) or supervised (trained). For example, the association of proprioceptive and exteroceptive stimuli can enable a Kohonen

[4] It would be more accurate to say that symbols should play no role since it has been noted that connectionist systems often fall back in the cognitivist paradigm by treating neural weights as a distributed symbolic representation [465].

neural network, controlling a robot arm, to learn hand-eye coordination so that the arm can reach for and track a visually-presented target. No *a priori* model of arm kinematics or image characteristics need be assumed [210, 298]. It has also been shown that a multilayered neural network with Hebb-like connectivity update rules can self-organize to produce feature-analyzing capabilities similar to those of the first few processing stages of the mammalian visual system (*e.g.* centre-surround cells and orientation-selective cells) [260].

Dynamical Systems Models

A dynamical system is an open dissipative non-linear non-equilibrium system: a system in the sense of a large number of interacting components with large number of degrees of freedom, dissipative in the sense that it diffuses energy (its phase space decreases in volume with time implying preferential sub-spaces), non-equilibrium in the sense that it is unable to maintain structure or function without external sources of energy, material, information (hence, open). The non-linearity is crucial: dissipation is not uniform and a small number of the system's degrees of freedom contribute to behaviour. These are termed *order parameters* (or *collective variables*). It is this ability to characterize a high-dimensional system with a low-dimensional model that is one of the features that distinguishes dynamical systems from connectionist systems.

Certain conditions must prevail before a system qualifies as a cognitive dynamical system. The components of the system must be related and interact with one another: any change in one component or aspect of the system must be dependent on and only on the states of the other components: 'they must be interactive and self contained' [465]. As we will see shortly, this is very redolent of the requirement for operational closure in enactive systems, the topic of the next section.

The key position of advocates of the dynamical systems approach to cognition is that motoric and perceptual systems are both dynamical systems, each of which self-organizes into meta-stable patterns of behaviour. 'Perceiving is not strictly speaking in the animal or an achievement of the animal's nervous system, but rather is a process in an animal-environment system' [219]. Perception-action coordination can also be characterized as a dynamical system.

Proponents of dynamical systems point to the fact that they provide one directly with many of the characteristics inherent in natural cognitive systems such as multi-stability, adaptability, pattern formation and recognition, intentionality, and learning. These are achieved purely as a function of dynamical laws and consequent self-organization through bifurcations and hysteresis in behavioural states, and a number of other dynamical properties such as intermittency (relative coordination).

Significantly, dynamical systems allow for the development of higher order cognitive functions in a straightforward manner, at least in principle. For example, intentionality — purposive or goal-directed behaviour — is achieved by the superposition of an intentional potential function on the intrinsic potential function [219]. Similarly, learning is viewed as the modification of already-existing behavioural patterns in the direction to be learned. It occurs in a historical context and the entire attractor layout (the phase-space configuration) of the dynamical system is modified. Thus, learning changes the whole

system as a new attractor is developed. Dynamical models can account for several non-trivial behaviours that require the integration of visual stimuli and motoric control. These include the perception of affordances, perception of time to contact, and figure-ground bi-stability [229, 147, 148, 485, 219].

The implications of dynamical models are many: as noted in [447], 'cognition is non-symbolic, nonrepresentational and all mental activity is emergent, situated, historical, and embodied'. It is also socially constructed, meaning that certain levels of cognition emerge from the dynamical interaction of between cognitive agents. Furthermore, dynamical cognitive systems are, of necessity, embodied. This requirement arises directly from the fact that the dynamics depend on self-organizing processes whereby the system differentiates itself as a distinct entity through its dynamical configuration and its interactive exploration of the environment.

One of key issues in dynamical systems is that cognitive processes are temporal processes that 'unfold' in real-time and synchronously with events in their environment. This strong requirement for synchronous development in the context of its environment again echoes the enactive systems approach set out in the next section. It is significant for two reasons. First, it places a strong limitation on the rate at which the ontogenic[5] learning of the cognitive system can proceed: it is constrained by the speed of coupling (*i.e.* the interaction) and not by the speed at which internal changes can occur [494]. Natural cognitive systems have a learning cycle measured in weeks, months, and years and, while it might be possible to collapse it into minutes and hours for an artificial system because of increases in the rate of internal adaptation and change, it cannot be reduced below the time-scale of the interaction (or structural coupling; see next section). If the system has to develop a cognitive ability that, *e.g.*, allows it to anticipate or predict action and events that occur over an extended time-scale (*e.g.* hours), it will take at least that length of time to learn. Second, taken together with the requirement for embodiment, we see that the consequent historical and situated nature of the systems means that one cannot short-circuit the ontogenic development. Specifically, you can't bootstrap an emergent dynamical system into an advanced state of learned behaviour.

Connectionist approaches differ from dynamical systems in a number of ways [219, 447, 465]. Suffice it here to note that the connectionist system is often defined by a general differential equation which is actually a schema that defines the operation of many (neural) units. That is, the differential equation applies to each unit and each unit is just a replication of a common type. This also means that there will be many independent state variables, one for each unit. Dynamical systems, on the other hand, are not made up of individual units all having the same defining equation and can't typically be so decomposed. Typically, there will be a small number of state variables that describe the behaviour of the system as a whole.

Enactive Systems Models

The last phase in the development of cognitive science comprises the study of enactive systems. Enaction is again an approach that is not based on the creation or use of representations. Significantly, cognitivism, by definition, involves a view of cognition

[5] Ontogeny is concerned with the development of the system over its lifetime.

that requires the representation of a given pre-determined world [471, 465]. Enaction adopts a critically different stance: cognition is a process whereby the issues that are important for the continued existence of the cognitive entity are brought out or enacted: co-determined by the entity as it interacts with the environment in which it is embedded. Thus, nothing is 'pre-given', and hence there is no need for representations. Instead there is an enactive interpretation: a context-based choosing of relevance. Enaction questions the entrenched assumption of scientific tradition that the world *as we experience it* is independent of the cognitive system ('the knower'). Instead, knower and known 'stand in relation to each other as mutual specification: they arise together' [471]. This type of statement is normally anathema to scientists as it seems to be positing a position of extreme subjectivism, the very antithesis of modern science. However, this is not what is intended at all. On the contrary, the enactive approach is an attempt to avoid the problems of both the realist (representationalist) and the solipsist (ungrounded subjectivism) positions. The only condition that is required of an enactive system is *effective action*: that it permits the continued integrity of the system involved. It is essentially a very neutral position, assuming only that there is the basis of order in the environment in which the cognitive system is embedded. From this point of view, cognition is exactly the process by which that order or some aspect of it is uncovered (or constructed) by the system. This immediately allows that there are different forms of reality (or relevance) that are dependent directly on the nature of the dynamics making up the cognitive system. Clearly, this is not a solipsist position of ungrounded subjectivism, but neither is it the commonly-held position of unique — representable — realism.

The enactive systems research agenda stretches back to the early 1970s in the work of computational biologists Maturana and Varela and has been taken up by others, even by some in the main-stream of classical AI [290, 291, 293, 470, 471, 494, 292].

The goal of enactive systems research is the complete treatment of the nature and emergence of autonomous, cognitive, social systems. It is founded on the concept of autopoiesis – literally *self-production* – whereby a system emerges as a coherent systemic entity, distinct from its environment, as a consequence of processes of self-organization. Three orders of system can be distinguished.

First-order autopoietic systems correspond to cellular entities that achieve a physical identity through structural coupling with their environment. As the system couples with its environment, it interacts with it in the sense that the environmental perturbations trigger structural changes 'that permit it to continue operating'.

Second-order systems are meta-cellular system that exhibit autopoiesis through operational closure in their organization; their identity is specified by a network of dynamic processes whose effects do not leave the network. Such systems also engage in structural coupling with their environment, this time through a nervous system that enables the association of many internal states with the different interactions in which the organism is involved.

Third-order systems exhibit (third-order) coupling between second-order (*i.e.* cognitive) systems, *i.e.* between distinct cognitive agents. These third-order couplings allow a recurrent (common) ontogenic drift in which the systems are reciprocally-coupled. The resultant structural adaptation – mutually shared by the coupled systems – gives rise to new phenomonological domains: language and a shared epistemology that reflects

(but not abstracts) the common medium in which they are coupled. Such systems are capable of three types of behaviour: (i) the instinctive behaviours that derive from the organizational principles that define it as an autopoietic system (and that emerge from the phylogenic evolution of the system), (ii) ontogenic behaviours that derive from the development of the system over its lifetime, and (iii) communicative behaviours that are a result of the third-order structural coupling between members of the society of entities. Linguistic behaviours are the intersection of ontogenic and communication behaviours and they facilitate the creation of a common understanding of the shared world that is the environment of the coupled systems. That is, *language is the emergent consequence of the third-order structural coupling of a socially-cohesive group of cognitive entities*.

The core of the enactive approach is that cognition is a process whereby a system identifies regularities as a consequence of co-determination of the cognitive activities themselves, such that the integrity of the system is preserved. In this approach, the nervous system (and a cognitive agent) does not 'pick up information' from the environment and therefore the popular metaphor of calling the brain an 'information processing device' is 'not only ambiguous but patently wrong' [292].

A key postulate of enactive systems is that reasoning, as we commonly conceive it, is the consequence of recursive application of the linguistic descriptive abilities (developed as a consequence of the consensual co-development of an epistemology in a society of phylogenically-identical agents) to the cognitive agent itself. This is significant: reasoning in this sense is a descriptive phenomena and is quite distinct from the mechanism (structural coupling and operational closure) by which the system/agent develops its cognitive and linguistic behaviours. Since language (and all inter-agent communication) is a manifestation of high-order cognition, specifically structural coupling and co-determination of consensual understanding amongst phylogenically-identical and ontogenically-compatible agents, reasoning is actually an artefact of higher-order social cognitive systems.

As with dynamical systems, enactive systems operate in synchronous real-time: cognitive processes must proceed synchronously with events in the systems environment as a direct consequence of the structural coupling and co-determination between system and environment. And, again, enactive systems are necessarily embodied systems. This is a direct consequence of the requirement of structural coupling of enactive systems. There is no semantic gap in emergent systems (connectionist, dynamical, or enactive): the system builds its own understanding as it develops and cognitive understanding emerges by co-determined exploratory learning. Overall, enactive systems offer a framework by which successively richer orders of cognitive capability can be achieved, from autonomy of a system through to the emergence of linguistic and communicative behaviours in societies of cognitive agents.

While the enactive systems agenda is very compelling, and is frequently referred to by researchers in, for example, developmental psychology, it hasn't achieved great acceptance in main-stream computational cognitive science and artificial intelligence. The main reason for this is that it is more a meta-theory than a theory *per se*: it is a philosophy of science but it doesn't offer any formal models by which cognitive systems can be either analysed or synthesized. However, it does have a great deal in common with the research agenda in dynamical systems which *is* a scientific theory but is perhaps

lacking the ability to prescribe how higher-order cognitive functions can be realized. The subsumption of the tenets of enactive systems into dynamical systems approaches may well provide the way forward for both communities, and for emergent approaches in general.

Finally, it is worth noting that the self-organizing constructivist approaches of both dynamical systems and enactive systems is bolstered by separate recent results which have shown that a biological organism's perception of its body and the dimensionality and geometry of the space in which it is embedded can be deduced (learned or discovered) by the organism from an analysis of the dependencies between motoric commands and consequent sensory data, without any knowledge or reference to an external model of the world or the physical structure of the organism [347, 348]. Thus, the perceived structure of reality could therefore be a consequence of an effort on the part of brains to account for the dependency between their inputs and their outputs in terms of a small number of parameters. Thus, there is in fact no need to rely on the classical idea of an *a priori* model of the external world that is mapped by the sensory apparatus to 'some kind of objective archetype'; that is, there is no need to rely on the cognitivist model of matters. On the contrary, the conceptions of space, geometry, and the world that the body distinguishes itself from arises from the sensorimotor interaction of the system. Furthermore, it is the analysis of the sensory consequences of motor commands that gives rise to these concepts. Significantly, the motor commands are *not* derived as a function of the sensory data. The primary issue is that sensory and motor information are treated simultaneously, and not from either a stimulus perspective or a motor control point of view.

2.3.3 Hybrid Models

Recently, some work has been done on approaches which combine aspects of the emergent systems and information processing & symbolic representational systems [152, 153, 155]. These hybrid approaches have their roots in strong criticism of the use of explicit programmer-based knowledge in the creation of artificially-intelligent systems [106] and in the development of active 'animate' perceptual systems [26] in which perception-action behaviours become the focus, rather than the perceptual abstraction of representations. Such systems still use representations and representational invariances but it has been argued that these representations can only be constructed by the system itself as it interacts with and explores the world. More recently, this approach has hardened even further with the work of Granlund who asserts that 'our conscious perception of the external world is in terms of the actions we can perform upon the objects around us' [152]. His is an approach of action-dependent vision in which objects should be represented as 'invariant combinations of percepts and responses where the invariances (which are not restricted to geometric properties) need to be learned through interaction rather than specified or programmed *a priori*'. Thus, a system's ability to interpret objects and the external world is dependent on its ability to flexibly interact with it and interaction is an organizing mechanism that drives a coherence of association between perception and action. There are two important consequences of this approach of action-dependent perception. First, one cannot have any meaningful access to the internal semantic representations, and second cognitive systems must be embodied (at least during the learning phase) [153]. For Granlund, action precedes perception and

Table 2.1. Attributes of different approaches to cognition (after [447] and [471])

Approaches to Cognition			
Cognitivist	**Connectionist**	**Dynamical**	**Enactive**
What is cognition?			
Symbolic computation: rule-based manipulation of symbols	The emergence of global states in a network of simple components	A history of activity that brings forth change and activity	Effective action: history of structural coupling that enacts (brings forth) a world
How does it work?			
Through any device that can manipulate symbols	Through local rules and changes in the connectivity of the elements	Through the self-organizing processes of interconnected sensorimotor subnetworks	Through a network of interconnected elements capable of structural changes
What does a good cognitive system do?			
Represent the stable truths of the real world	Develop emergent properties that yield stable solutions to tasks	Become an active and adaptive part of an ongoing and continually changing world	Become a part of an existing world of meaning (ontogeny) or shape a new one (phylogeny)

'cognitive systems need to acquire information about the external world through learning or association' ... 'Ultimately, a key issue is to achieve behavioural plasticity, *i.e.*, the ability of an embodied system to learn to do a task it was not explicitly designed for.' Thus, action-dependent systems are in many ways consistent with emergent systems while still exploiting programmer-centred representations. The identification of these representations (and processes that generate them) could be construed as being equivalent to the phylogenic specification of an emergent system, an issue that has yet to be resolved.

To summarize, Table 2.1 (after [447] and [471]) contrasts the four approaches (the cognitivist and the three emergent approaches) under three broad headings: *What is cognition? How does it work?* and *What does a good cognitive system do?*

2.4 The Space of Cognitive Vision: Knowledge & Memory, Deliberation, and Learning

It is clear from the previous section that there are several distinct and fundamentally different approaches to cognition. If we are to tackle the problem of cognitive vision in a consistent and coherent manner, we need to embed these approaches in a paradigm-independent framework that spans the entire spectrum of approaches, makes explicit their position in the spectrum, but also that highlights technical issues of mutual relevance. It is proposed that, leaving aside the four levels of generic visual functionality identified in Section 2.2, the three concerns of knowledge & memory, deliberation, and learning effectively span the space of cognitive vision.

2.4.1 Knowledge & Memory

Knowledge – its acquisition, retention, and use to deliberate about tasks and situations, and to achieve goals – is the central issue in cognitive systems. Two questions arise: where does the knowledge come from, and what relationship does it imply between the cognitive agent and its environment?

At one end of the spectrum, we have the information processing representational approach: knowledge is provided in symbolic form by human designers and possibly refined through learning, either off-line or on-line. These systems are often quite brittle: because their representations are based on the representational and processing axioms of an external designer, they often fail when their domain of discourse strays far from the domain for which they were designed.

Further along this spectrum, we have hybrid systems that use generic representational schemes (*e.g.* sub-space methods such as ISA, ICA, PCA) that are populated entirely by the agent through incremental unsupervised learning, either off-line or on-line. Deliberative cognitive behaviour is designed in, since the designer controls the link between the perceptual-motor skills and the knowledge representations. On the other hand, the system learns the actual knowledge for itself. Thus, its cognitive behaviour is a function of agent-specific and agent-acquired knowledge and pre-programmed perception-reasoning-action strategies. These types of approaches probably represent the state of the art in modern cognitive computer vision systems, both in terms of engineering maturity and functional capability.

Further along again, we move into the emergent systems space populated by, *inter alia*, connectionist approaches and dynamical systems approaches. These have no 'representations' in the sense that there are no symbols that refer to world-based objects and situations, but they do have states that encapsulate knowledge derived by the system in the historical context of its ontogenic development. In collectives or societies of cognitive agents with linguistic and communicative abilities, it is possible for a cognitive agent to describe its visual environment. These descriptions are effectively symbolic representations but they are a consequence of social cognitive behaviour, *not* the mechanism by which cognition emerges (in contradistinction to the cognitivist approach) .

2.4.2 Deliberation

Briefly, one can distinguish between 'non-thinking' cognitive agents (with reflexive autonomous behaviour), 'thinking' cognitive agents (with anticipatory adaptive behaviour), and 'thinking about thinking' cognitive agents (with the ability to explicitly reflect on their cognitive processes). At the information processing representationalist cognitivist end of the spectrum, deliberation is identical with symbolic reasoning. Deliberation *is* the operation on symbolic information representations by processes of logical reasoning. Thus thinking and thinking about thinking can be viewed as some formal process of symbol manipulation. This is exactly the cognitivist hypothesis.

On the other hand, emergent systems take a fundamentally different view. Non-thinking agents don't deliberate; their reflexive action is a consequence of their phylogenic configuration (*i.e.* their innate perceptual-motor behaviours). Thinking cognitive agents engage in weak deliberation: their anticipatory and adaptive behaviour

emerges because of the evolution of its state-space as a consequence of its real-time co-development with its environment (*i.e.* ontogenic learning). Strong deliberation ("if I do this, then this will happen") arises in agents that can think about thinking (or think about thoughts, including possible actions). The emergence of this possibility arises when a society of cognitive agents with linguistic skill co-evolve and develop a common epistemological understanding of their world; the linguistic skills and the epistemological constructs give rise to the possibility of descriptions of observations, which when turned inward recursively on the cognitive agent, facilitates thinking about thinking, or exactly the reasoned discourse that we commonly identify (incorrectly) with cognition: such reasoned discourse – thinking – is an artifact of the systems socio-linguistic ontogenic development in a society of such agents and is not a causal process in the emergence of cognition.

2.4.3 Learning

The issue of learning is central to cognitive systems. Unfortunately, it is also an ill-posed concept in the sense that what learning means depends on which view of cognition you hold. Different approaches to cognition imply different assumptions and different processes. That said, one can define learning in a general way that is not specific to a given paradigm as the adaptive change in the state of the cognitive system as a result of a temporal process whereby the system uses experience to improve its performance at some defined behaviour.

For systems built around the cognitivist information processing representational view of cognition, learning typically involves the alteration of some representational data-structure. Learning can be accomplished in many ways using, for example, probabilistic (e.g. Bayesian) inference techniques or non-probabilistic approaches such as Dempster-Shafer transferrable belief models. However, there may be other approaches and the literature on machine learning is vast and growing.

For emergent systems, learning is accomplished in a manner that depends on the technique being deployed. In dynamical systems, such those advocated by developmental psychologists, the entire attractor layout of the dynamical system is modified, even in a single instance of learning. That is, learning changes the whole system. Learning moulds the intrinsic dynamics. Once learning is achieved, the memorized or learned pattern constitutes a new attractor. It is significant that learning is a specific modification of already existing behavioural patterns. This implies that learning occurs in the historical context of an individual and is dependent on the history of experiences of the learner. That is, learning is situated, historical, and developmental.

In the same way, for enactive systems, as the system couples with its environment, it interacts with it in the sense that the environmental perturbations trigger structural changes that permit it to continue operating. These changes are governed by the operational closure of autopoeitic organization. An observer of the system would describe the structural changes that occur in the (nervous system) of the autopoietic entity as a learning process; however, from the perspective of the entity itself, there is just the ongoing natural structural drift: no information is picked up from the environment, but the system evolves in co-determination with the environment as it interacts with it.

Connectionist approaches distinguish several types of learning: 'learning-by-doing' or unsupervised associative learning, supervised learning (intelligent teaching with explicit input-output exemplars), unsupervised learning (clustering algorithms), and reinforcement learning (use of a cost function). 'Learning-by-doing' can be traced to Piaget, who founded the constructivist school of cognitive development in the 1940's [298]. According to this view of constructivism, knowledge is not imparted *a priori* but is discovered and constructed by a child though active manipulation of the environment.

In general, for emergent approaches, one can view learning as the ontogenic development of the cognitive agent, irrespective of how that development is achieved. Certainly, it is dependent to some extent on the phylogenic make-up of the system and the innate (reflexive) behaviours.

Winograd and Flores [494] capture the essence of developmental emergent learning very succinctly:

> 'Learning is not a process of accumulation of representations of the environment; it is a continuous process of transformation of behaviour through continuous change in the capacity of the nervous system to synthesize it. Recall does not depend on the indefinite retention of a structural invariant that represents an entity (an idea, image, or symbol), but on the functional ability of the system to create, when certain recurrent conditions are given, a behaviour that satisfies the recurrent demands or that the observer would class as a reenacting of a previous one'.

2.5 Some Implications of the Different Approaches

Two of the most important questions in cognitive vision and cognition are:

1. Can you engineer knowledge and understanding into a system, providing it with the semantic information required to operate adaptively and achieve robust and innovative goal-directed behaviour?
2. Does a cognitive system necessarily have to be embodied (in the sense of being a physical mobile exploratory agent capable of manipulative and social interaction with its environment, including other agents)?

It is clear that the answers to these questions will be different depending on who you ask. The person who subscribes to the cognitivist approach will answer respectively, 'yes, you can engineer knowledge into a system' and 'no, it doesn't need to be embodied'. Conversely, a person working with emergent systems (including connectionist systems, dynamical systems, enactive systems, and hybrid approaches) will answer respectively, 'no, you can't engineer knowledge into a system' and 'yes, it does need to be embodied'.

For those adopting the cognitivist approach, knowledge and world (scene) representations are primary: they allow explicit description of situations and structure, and of spatio-temporal behaviours. Typically, they deal with categories and categorization by focussing on form rather than function. Their tools include qualitative descriptions and conceptual knowledge, reasoning (over representations), inference over partial information, and experience/expectation driven interpretation. Interaction with humans

is achievable exactly because the information representations are consistent with the (human) system designer's pre-conceptions. Motoric control is optional because these systems don't necessarily have to be embodied: knowledge, framed in the concepts of the designer, can be transplanted in and doesn't have to be developed by the system itself through exploratory investigation of the environment. Hybrid cognitivist systems, however, do exploit learning (on-line and unsupervised) to augment or even supplant the *a priori* designed-in knowledge and thereby achieve a greater degree of adaptiveness, reconfigurability, and robustness. Consequently, embodiment, while not strictly necessary in this approach, does offer an additional degree of freedom in the research agenda. In this case, agent intentions (and intentional models) require robust perception and learning to build the models, and robust reasoning capabilities to produce sensible communication with humans and/or its motoric interface. Ultimately, knowledge is used both to pre-configure a system for operation in a particular domain and to provide it with the constraints required to render the problem tractable.

It has been argued, however, that this approach to cognition is inherently limited. Since 'the essence of intelligence [*qua* cognition] is to act appropriately when there is no simple pre-definition of the problem or the space of states in which to search for a solution' [494] and since, in the cognitivist approach, the programmer sets up the system with a systematic correspondence between representations and the programmer's entities, the cognitive vision system is effectively blinded because the programmer's success as a cognitive entity in his or her own right goes far beyond the reflective conscious experience of things in the world. Thus, according to this argument, the cognitivist approach is inherently incapable of exhibiting fully-fledged cognitive capabilities, although as pointed out several times, it can of course be used to develop useful cognitive capabilities in circumstances which are well-defined and well-bounded. In this case, cognition requires either expectation-driven constraints or *a priori* programmer knowledge to render the cognitive vision problem tractable.

There are also several consequences of adopting the emergent systems approach. Phylogenic development (the evolution of the system configuration from generation to generation) gives rise to the system's innate reflexive capabilities. Ontogenic development (the adaptation and learning of the system over its lifetime) gives rise to the cognitive capabilities that we seek. Since we don't have the luxury of having evolutionary timescales to allow phylogenic emergence of a cognitive system, we must somehow identify a minimal phylogenic state of the system. Operationally, this means that we must identify and effect a mechanism for the minimal reflexive behaviours required for subsequent ontogenic development of cognitive behaviour. For dynamical systems, this is equivalent to the identification of the collective variables and the control parameters. A major research issue is how to accomplish this without falling back into conventional cognitivism: system identification based on representations derived from external observers.

A prerequisite condition for the emergence of cognitive systems (either connectionist, dynamical, or enactive) is operational (organizational) closure: the system components must be related and interact with one another and any change in one component or aspect of the system must be dependent on and only on the states of the other components. This provides some boundary conditions on the identification of admissible architectures.

The descriptions of a cognitive system are a function of its interaction with the environment and the richness of its action interface. This implies that a cognitive system must have non-trivial manipulative and exploratory capabilities.

Cognitive systems can only engage in linguistic communication with agents that have shared the same ontogenic developmental learning experience. This poses a significant hurdle for proponents of emergent systems since the overt purpose of an artificial cognitive agent is to engage in behaviours that are relevant to human users.

The requirement of real-time synchronous system-environment coupling implies strong constraints on the rate at which ontogenic learning can proceed and necessitates historical, situated, and embodied development that can't be short-circuited or interrupted. Again, this is a difficult issue for empirical research in cognitive systems: it would be more than a little frustrating to develop a cognitive system after many days of developmental learning only to have it disappear because of a power glitch.

2.6 Conclusion

Broadly speaking, there are essentially two approaches to cognition:

1. The cognitivist symbolic information processing representational approach;
2. The emergent systems approach (connectionism, dynamical systems, enactive systems).

The one thing that is common to both is the issue of knowledge and thinking. However, each approach takes a very different stance on knowledge.

The cognitivist approach:

- takes a predominantly static view of knowledge, represented by symbol systems that *refer* (in a bijective and isomorphic sense) to the physical reality that is external to the cognitive agent;
- invokes processes of reasoning on the representations (that have been provided by the perceptual apparatus);
- subsequently plans actions in order to achieve programmed goals;
- and can be best epitomized by the classical perception-reasoning-action cycle.

The emergent systems approach:

- takes a predominantly dynamic or process view of knowledge, and views it more as a collection of abilities that encapsulate 'how to do' things;
- is therefore subservient to the cognitive agent and dependent on the agent and its environmental context;
- embraces both short time-scale (reflexive and adaptive) behaviours and longer time-scale (deliberative and cognitive) behaviours, with behaviours being predominantly characterized by perceptual-motor skills;
- is focussed on the emergence (or appearance) of cognition through the co-development of the agent and its environment in real-time.

There is one other crucial difference between the two approaches. In the cognitivist symbolic information processing representational paradigm, perceptual capacities are configured as a consequence of observations, descriptions, and models of an external designer (*i.e.* they are fundamentally based in the frame-of-reference of the designer). In the emergent systems paradigm, the perceptual capacities are a consequence of an historic enactive embodied development and, consequently, are dependent on the richness of the motoric interface of the cognitive agent with its world. That is, the action space defines the perceptual space and thus is fundamentally based in the frame-of-reference of the agent. A central tenet of the enactive emergent approaches is that true cognition can only be created in a developmental agent-centred manner, through interaction, learning, and co-development with the environment.

So much for differences. There is, however, also some common ground. In our attempt at a universal definition of cognitive vision, we identified three categories of requirements for a cognitive vision system:

1. the four generic visual capabilities of detection, localization, recognition, and understanding;
2. the non-functional attributes of purposive goal-directed behaviour;
3. the three knowledge-based faculties of learning, memory, and deliberation that we subsequently used to span the space of cognitive vision.

Regarding visual capability, we noted that all emergent approaches face the problem that the evolution of a cognitive vision system *ab initio* is simply not feasible and, consequently, there is a need to identify the minimal phylogenic structure of the system, *i.e.* a minimal set of visual (and motoric) capabilities. Thus, both emergent cognitive vision and cognitivist cognitive vision, as well as the hybrid approaches, need solid robust image processing to facilitate the four generic visual capabilities and it is likely that visual processes are transferrable between paradigms.

Although the non-functional attributes are inevitably going to be paradigm-dependent, it remains to be seen how much each paradigm overlaps in the three dimensions of learning, memory, and deliberation. One thing is certain: there is a much better chance of spotting overlaps and common ground if everyone is using the same framework. The hope is that this paper will go some way towards creating that framework.

Acknowledgments

The discussions at Dagstughl seminar 03441 [73] on cognitive vision systems were pivotal in shaping the ideas presented in this paper, alerting us to the necessity of keeping open a broad research agenda without losing the focus on well-engineered computer vision systems.

Discussions with Giulio Sandini and Giorgio Metta, University of Genova, were also extremely helpful in sifting out the key issues and showing a way to actually achieve an inclusive and realistic framework that is, hopefully, relevant to all.

This work was funded by the European Commission as part of the European research network for cognitive computer vision systems — *ECVision* — under the Information Society Technologies (IST) programme, project IST-2001-35454.